Diverse Interdisciplinary Research Teams and Secure Access To HIPAA Protected Data

Dr. D. Gardner, Dr. M-A. Demuynck, (Mathematics and Computer Science), Dr. M. Tietze (Nursing)

PAST

Context

Ensuring the protection and privacy of health data while still facilitating access for a diverse team of researchers can be a difficult task, given the financial and personnel resource constraints in place at many organizations. For small universities with limited budgets, creating HIPAA compliant research opportunities can be particularly challenging. This project sought to develop a research data environment for a faculty research team, whose members are spread across three campuses and multiple disciplines, in their effort to analyze mental health data from North Texas area hospitals obtained from the DFWHC Foundation Warehouse.

Obstacles

The data set consisted of over 870,000 records containing over 120 fields and was initially available as a flat file housed on an encrypted Windows network share and accessible via the Excel spreadsheet program.

Due to load on the file server, which contained many private shares for use across the organization and the large size of the single file, file load and access times were excruciatingly slow. This issue was further compounded when sorting, cleaning, reorganization, and analysis was attempted on the data.

To support the analysis needs of multiple individuals from different machines a dedicated HIPAA compliant data server running a relational database was needed to house, protect, and support more robust and complex analysis of the data.

Unfortunately, there were not funds available to support the purchase of physical hardware, purchase commercial encrypted data software solutions, or to hire support staff to configure and administer the data server. Additionally, when doing a broad analysis of the data, it was found that the data in the current form could not easily be exported into a UTF-8 encoded CSV for importation into a database because of restrictions in Excel and artifacts such as comma, tabs, and null fields within the data.

PRESENT

Data Server Information

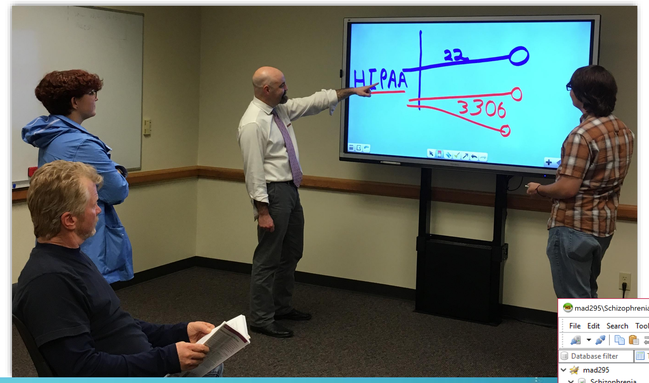
- Perl and Python scripts in conjunction with open source modules such as Pandas library developed for encoding conversion, data cleaning, and data migration
- HIPAA virtual machine *farm* on isolated section of internally accessible network/domain
- Debian Linux based virtual machine
- MariaDB™, fork of and drop-in replacement for MySQL database server software
- 256 bit encryption keys and certificates for server and clients generated
- 256 bit encryption for all data within database using open source plugins
- Secondary encrypted backup database of original data set
- Option for database administrator in addition to system administrator to create and manage additional tables as requested by researchers
- Remote database allows for optional local frontends such as MySQL Workbench™, HeidiSQL™, SAS™, etc.
- Identifying HIPAA requirements for data privacy and protection for team members unfamiliar with rules and regulations

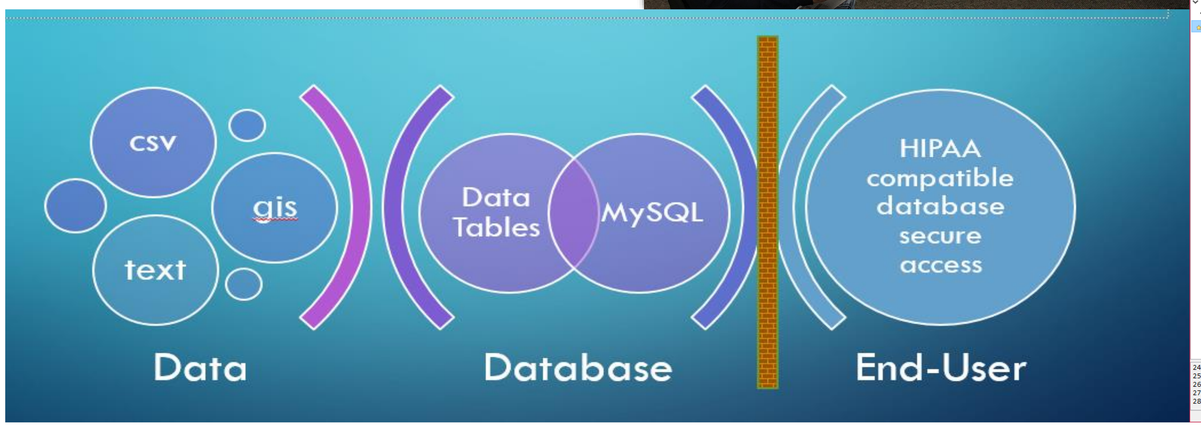
User Training

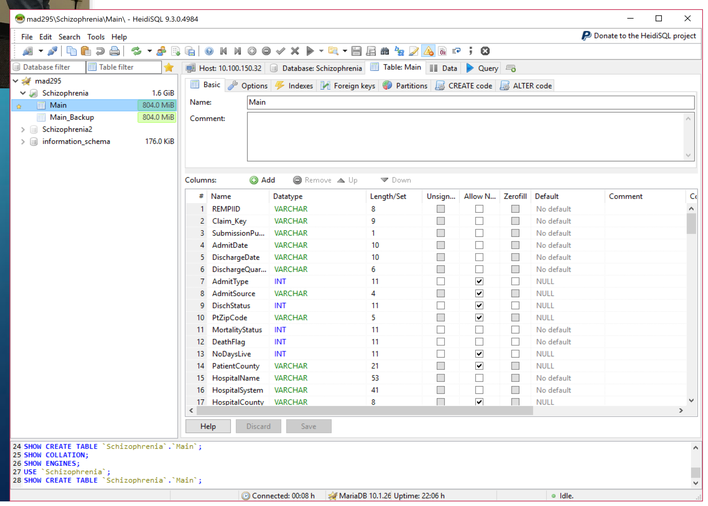
- Prepare user training materials
- Hold user training sessions

Access Control

- Combination of network firewalls, packet filtering, and proxies used by IT to control access to farm and servers in general on network
- On Server iptables firewall allowing traffic across port 3306 from specific IPs
- Port 22 SSH access only allowed for system admin from admin workstation
- Static IPs for researcher workstations and lab workstations
- VPN access to workstations upon request
- TLS based 256 bit encryption required for all logins across port 3306
- Database user accounts tied to personal workstation IPs
- Logins from all researcher database accounts on lab workstations
- Use of graduated roles to manage account and database privileges instead of per user privilege management







FUTURE

Database Improvements

- Additional data sanitization
- Splitting up of main table into appropriate smaller tables for specific subsets of the larger data set along with the creation of primary keys and relationships as needed
- Automatic backup of working database on a separate VM
- Creation of a separate VM to act as a key server to support rolling ciphers.

Access Control

- Client certificates added and required for all client workstations
- Creation of plugin or script to aid researchers in configuration and setup of preferred frontend

Automation

- Server and client key/certificate generation and organization scripts
- Automatic plugin and database server configuration scripts to enable and configure encryption for created databases within the server software
- Scripts to generate appropriate roles and accounts based on provided list of researchers/users
- Scripts to automate the installation of required packages and other platform configuration
- Master script to automate individual component scripts